(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－184891

(43)公開日　平成11年(1999)７月９日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/403	３４０Ａ
	13/00	３５１		13/00	３５１Ｇ
				15/40	３１０Ｆ
					３１０Ｃ
				15/403	３４０Ｂ

審査請求　未請求　請求項の数４　ＦＤ　（全 13 頁）

(21)出願番号　特願平9－364536

(22)出願日　平成９年(1997)12月18日

(71)出願人　000005496
富士ゼロックス株式会社
東京都港区赤坂二丁目17番22号

(72)発明者　広瀬　真
神奈川県足柄上郡中井町境430　グリーンテクなかい　富士ゼロックス株式会社内

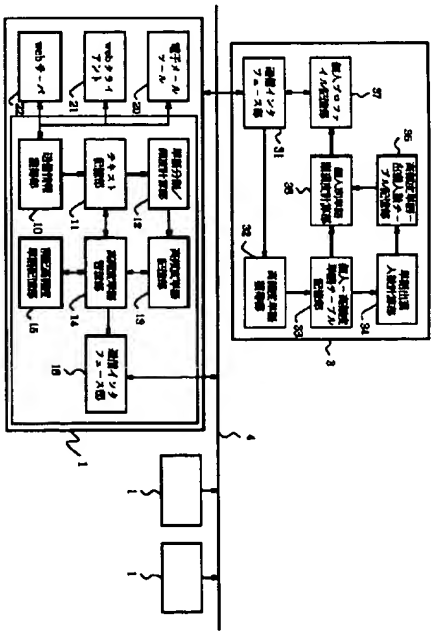
(74)代理人　弁理士　守山　辰雄

(54)【発明の名称】　個人プロファイル管理装置

(57)【要約】

【課題】　個人の専門性や興味等を精度よく表した個人プロファイルを、テキストの内容の他者への開示を制限して生成する。

【解決手段】　個人プロファイル管理装置は、複数のクライアントシステム１と少なくとも１つのサーバシステム３とを有し、クライアントシステムを利用する個人に関する個人プロファイルを作成する。クライアントシステム１では、送信情報獲得部１０により個人が他者に送信するテキストを獲得し、単語分割／頻度計算部１２が当該テキストから単語を抜き出すとともに出現回数を計数して、単語データを生成する。サーバシステム３では、高頻度単語獲得部３２により単語データを複数のクライアントシステムから受信し、単語出現人数計算部３４が複数の単語データから複数の個人間で同一の単語が出現する人数を単語人数データとして計算し、単語顕現度計算部３６が単語人数データに基づいて個人毎の単語データを補正して、他者に対する各個人の単語の相対的な顕現性の度合いを個人プロファイルとして定める。



1

【特許請求の範囲】
【請求項１】　複数のクライアントシステムと当該クライアントシステム間の通信を管理するサーバシステムとを有し、当該クライアントシステムを利用する個人に関する情報を管理する個人プロファイル管理装置であって、
前記クライアントシステムは、
当該クライアントシステムを利用する個人が他者に送信するテキストを獲得する送信情報獲得手段と、
前記テキストから単語を抜き出すとともに当該単語の出現回数を計数して、当該単語に出現頻度を対応付けた単語データを生成する単語分割／頻度計算手段と、を有し、
前記サーバシステムは、
前記単語データを複数のクライアントシステムから受信する高頻度単語獲得手段と、
前記受信した複数の単語データから複数の個人間で同一の単語が出現する人数を単語人数データとして計算する単語出現人数計算手段と、
前記単語人数データに基づいて個人毎の単語データを補正して、他者に対する各個人の単語の相対的な顕現性の度合いを個人プロファイルとして定める単語顕現度決定手段と、を有することを特徴とする個人プロファイル管理装置。
【請求項２】　複数のクライアントシステムを有し、当該クライアントシステムを利用する各個人に関する情報を管理する個人プロファイル管理装置であって、
前記クライアントは、
当該クライアントシステムを利用する個人が他者に送信するテキストを獲得する送信情報獲得手段と、
他のクライアントシステムを利用する他者から受信したテキストを獲得する受信情報獲得手段と、
前記送信するテキストと前記受信したテキストとから単語を抜き出すとともに当該単語の出現回数を計数して、当該単語に出現頻度を対応付けた単語データを個人毎に生成する単語分割／頻度計算手段と、
前記複数の単語データから複数の個人間で同一の単語が出現する人数を単語人数データとして計算する単語出現人数計算手段と、
前記単語人数データに基づいて個人毎の単語データを補正して、他者に対する各個人の単語の相対的な顕現性の度合いを個人プロファイルとして定める単語顕現度決定手段と、を有することを特徴とする個人プロファイル管理装置。
【請求項３】　請求項１または請求項２に記載の個人プロファイル管理装置において、
前記単語分割／頻度計算手段は所定の条件を満たす一定以上の出現頻度の単語について単語データを生成することを特徴とする個人プロファイル管理装置。
【請求項４】　請求項１乃至請求項３のいずれか１項に記載の個人プロファイル管理装置において、
前記クライアントシステムは、
前記単語分割／頻度計算手段が過去に生成した単語データを保持する記憶手段と、
前記単語分割／頻度計算手段が生成した単語データと前記記憶手段に保持された過去の単語データとを１つの単語データに合成する単語管理手段と、
をさらに有することを特徴とする個人プロファイル管理装置。

2

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、個人の専門領域や興味等を知るための情報として重要な単語をプロファイルとしてテキスト中から抽出する個人プロファイル管理装置に関し、特に、複数の個人間で各個人に顕現度の高い単語を抽出する個人プロファイル管理装置に関する。
【０００２】
【従来の技術】電子化された大量の情報の中から自分にとって有用な情報のみを選択的に参照するために、個人の興味を登録した個人プロファイルを有する情報フィルタリング装置が提案されている。また、多数の人物の中から自分にとって有用な情報を持つ人物のみと選択的にコミュニケーションを行うために、個人の興味を登録した個人プロファイルを有する個人プロファイル検索装置が提案されている。このような個人プロファイルは、個人の興味等を特徴付けるために属性名と属性値の組み、あるいは、複数のフリーキーワード等によって構成されている。
【０００３】ここで、これらの個人プロファイルは、本人が自己について記述した場合には、本当の専門性や興味を表現していないことや、興味の時間的な変化に合わせて個人プロファイルも更新しなければならない等の問題点があり、これに対処するために、個人プロファイルを自動的に抽出する技術が提案されている。この抽出技術では、例えば特開平８－２３５０８８号公報に開示されるように、送受信されるテキスト情報から個人の興味を表現する複数の単語を抜き出し、該個人が該テキスト情報に対して行った処理の頻度情報などを活用して、プロファイルに含まれる個々の項目の優先度を該個人の興味の実態に合致するように調整している。
【０００４】しかしながら、個人の興味を的確に表現しているだけでは、情報フィルタリング装置あるいは個人プロファイル検索装置の手段として利用する場合、他者との相対的な関係に起因する問題点があった。例えば、或る個人の興味を的確に表現する単語群の内の上位に位置する単語が、「情報」や「コンピュータ」等であった場合、専門分野を特定していない集団においては十分に個人プロファイルとして機能するが、情報やコンピュータに興味を持っている集団においては、他の多くの人々の個人プロファイル中にも同じ単語が出現するために、

「情報」「コンピュータ」は個人を特徴付ける単語にはならない。したがって、個人の興味を的確に表現しているだけでは、集団における個人の特徴を表現するために最適なプロファイルを抽出することができなかった。

【０００５】なお、個人プロファイルの抽出を目的とした技術ではないが、他のテキストとの相対的な関係を考慮したテキストに対するキーワード抽出技術が知られている。例えば、特開平２－２４４２７４号公報に開示されるように、或る単語について、一つのテキスト内に出現する比率（単語の出現回数÷単語の総数）と、或る領域のテキスト集合中に出現する比率（領域内の単語の出現回数÷領域内の単語の総数）との比率の大小を考慮して、キーワードを選択する。

【０００６】また、ｔｆｉｄｆ理論（G. Salton & C. B*

*uckley, "Term Weighting Approaches in Automatic Text Retrieval", Department of Computer Science, Cornell University, 87-881, November, 1987）のようなキーワードの重み付けの方法が知られている。このｔｆｉｄｆ理論においては、ｔｆｉｋをテキストＤｉにおけるキーワードＴｋの出現回数、Ｎを全テキスト数、ｎｋを全テキストの内のキーワードＴｋを含むテキスト数とするとき、テキストＤｉにおけるキーワードＴｋの重みｗｉｋを次式（１）で決定する。このことにより、テキスト中の出現回数が多く、かつ、他のテキスト中の出現比率が低い単語をキーワードとして選択することが可能になる。

【０００７】

【数１】

$$W_{ik} = \{tf_{ik} \times \log (N \div n_k)\} \div \{\sum_{k=1}^{t} (tf_{ik})^2 \times (\log (N \div n_k))^2\}^{0.5} \quad \cdots\cdots (1)$$

【０００８】

【発明が解決しようとする課題】前述のキーワード抽出技術は、個人プロファイル抽出を目的に考案された方法ではないが、他のテキストとの相対的な関係を利用するという考え方は、前述の個人プロファイル抽出技術の欠点を克服できる可能性がある。つまり、キーワード抽出技術における１テキストを、１人の個人に係わるテキスト群とし、他の一群のテキスト集合を、他の一群の人物集合に係わる全テキスト集合とみなすことで、他者のテキスト群にはあまり表れないが、その個人のテキスト群には頻繁に表れる単語の重みを大きくすることが実現可能であるとも考えられる。

【０００９】しかしながら、本来、テキスト群中の各々のテキストに適切なキーワードを付与するために考案されたキーワード抽出技術を、人物の集団の中の個々の人物に適切なプロファイル情報を作成する目的に応用しようとすると以下に述べるような問題点があった。なお、以下では、個人プロファイルを抽出するために好適な特性を持つ対象テキストとして、個人が送受信した電子メールテキスト群を例に説明する。

【００１０】「内容の他者への開示制限」個人プロファイルを抽出する対象となるテキストは、内容を他者に開示されることを制限したいという特徴がある。例えば、電子メールの内容は私信に相当し、他者への内容の開示は可能な限り避けたいという要求がある。しかしながら、ｔｆｉｄｆなどに代表されるキーワード抽出技術は、キーワードの重みを計算する度に、全対象テキストの内容のスキャン処理を行う必要のあるアルゴリズムである。したがって、個人プロファイルの抽出処理が他者の管理下にある装置で行われる場合は、計算の都度、電子メールテキスト群の全文を他者の管理下に預けなければならないという問題点があった。一方、個人プロファイルの抽出処理を自分の管理下にある装置で行う場合には、逆に、他者の電子メールテキスト群の全文を預かる必要がある。また、個人プロファイルの抽出処理を信頼できる第三者の管理下にある装置で行う場合でも、全文を預けなければならず、利用者の心理的な負担の根本的な解決にはなっていない。

【００１１】したがって、個人プロファイル抽出の対象となるテキスト群に、従来のキーワード抽出処理をそのまま応用すると、内容の他者への開示を制限したいという要求に反してしまう。そのため、従来のキーワード抽出処理を、各自の管理下にある装置上での処理と、第三者の管理下にある装置での処理とに、適切に分散させることが必要になってくるが、その実現方法は従来においては何ら考慮されていない。

【００１２】「対象文書の非保存性」計算機の記憶手段の容量は有限であるが、個人プロファイルを抽出する対象となるテキスト群は一過性のテキストが多い。例えば、パソコン通信サービスでは、一定の期間が経過すると送信済みの電子メールのテキストは削除され、記憶容量の有効利用を図っている。しかしながら、前述のキーワード抽出技術は、キーワードの重みを計算する度に、全対象テキストの内容のスキャン処理を行う必要のあるアルゴリズムである。したがって、個人の管理下であるか否かを問わず、重みの計算以外には利用価値のないテキスト群を記憶しておく必要がある。

【００１３】したがって、個人プロファイル抽出の対象となるテキスト群に、従来のキーワード抽出処理をそのまま応用すると、不要なテキストが記憶容量を圧迫するという問題が生じる。この問題の解決のために、個人プロファイル抽出の処理の中に、情報を圧縮して保存しておくアルゴリズムを導入する必要があるが、その実現方法は従来においては何ら考慮されていない。

【００１４】「変化への追従性」個人の専門性や興味は時々刻々変化をする。前述のキーワード抽出技術では、個々のテキストの作成日時情報などを参照しながら、最新のテキスト群だけを選択すること等で対処可能な問題である。しかしながら、専門性や興味の変化に追従し続けるという課題は、上記の他者への開示制限と非保存性の問題と同時に解決する必要がある。処理の分散と情報の圧縮を安易に行うことは、変化への追従性を確保するために重要な、対象テキストの再選択を困難なものにしてしまう。したがって、他者への開示制限と非保存性の問題に対する処理の分散と情報の圧縮という解決は、個人情報の変化への追従性を保持した解決でなければならないが、その実現方法は従来においては何ら考慮されていない。

【００１５】本発明は上記従来の事情に鑑みなされたもので、個人の専門性や興味等を精度よく表した個人プロファイルを生成することができる個人プロファイル管理装置を提供することを目的とする。さらに、本発明は、電子メールテキスト等の内容の他者への開示を制限して、このような個人プロファイルを生成することができる個人プロファイル管理装置を提供することを目的とする。

【００１６】

【課題を解決するための手段】本発明に係る個人プロファイル管理装置は、複数のクライアントシステムと少なくとも１つのサーバシステムとを有し、当該クライアントシステムを利用する個人に関する情報を管理する個人プロファイル管理装置として構成される。そして、このクライアントシステムでは、送信情報獲得手段により当該クライアントシステムを利用する個人が他者に送信する（すなわち、他者へ送信した、あるいは、他者が取得可能な状態とした）テキストを獲得し、単語分割／頻度計算手段が当該テキストから単語を抜き出すとともに当該単語の出現回数を計数して、当該単語に出現頻度を対応付けた単語データを生成する。一方、サーバシステムでは、高頻度単語獲得手段により単語データを複数のクライアントシステムから受信し、単語出現人数計算手段が当該受信した複数の単語データから複数の個人間で同一の単語が出現する人数を単語人数データとして計算し、さらに、単語顕現度決定手段が当該単語人数データに基づいて個人毎の単語データを補正して、他者に対する各個人の単語の相対的な顕現性の度合いを個人プロファイルとして定める。

【００１７】したがって、テキストは各個人が利用するクライアントシステムで処理されるため、電子メールテキスト等のように内容の他者への開示を制限したいという要求を満すことができ、しかも、他者に対する各個人の単語の相対的な顕現度を加味して個人プロファイルを定めるため、各個人の特徴をよく表した個人プロファイルを生成することができる。

【００１８】また、本発明に係る個人プロファイル管理装置は、複数のクライアントシステムを有し、当該クライアントシステムを利用する各個人に関する情報を管理する個人プロファイル管理装置として構成される。そして、このクライアントシステムでは、送信情報獲得手段により当該クライアントシステムを利用する個人が他者に送信するテキストを獲得し、また、受信情報獲得手段により他のクライアントシステムを利用する他者から受信したテキストを獲得し、これらテキストから単語分割／頻度計算手段が単語を抜き出すとともに当該単語の出現回数を計数して、当該単語に出現頻度を対応付けた単語データを個人毎に生成する。そして、このクライアントシステムでは、単語出現人数計算手段が複数の単語データから複数の個人間で同一の単語が出現する人数を単語人数データとして計算し、この単語人数データに基づいて個人毎の単語データを補正して、他者に対する各個人の単語の相対的な顕現性の度合いを個人プロファイルとして定める。

【００１９】したがって、テキストは各個人が利用するクライアントシステムおよび送信先に指定した特定のクライアントシステムで処理されるため、電子メールテキスト等のように内容の他者への開示を制限したいという要求を満すことができ、しかも、他者に対する各個人の単語の相対的な顕現度を加味して個人プロファイルを定めるため、各個人の特徴をよく表した個人プロファイルを生成することができる。

【００２０】また、本発明に係る個人プロファイル管理装置では、単語分割／頻度計算手段は所定の条件を満たす一定以上の出現頻度の単語について単語データを生成し、これによって、必要性の低い単語についての処理を回避して処理負担を軽減するとともに処理に利用する記憶容量の低減化を実現する。また、本発明に係る個人プロファイル管理装置では、クライアントシステムは、単語分割／頻度計算手段が過去に生成した単語データを保持する記憶手段と、単語分割／頻度計算手段が生成した単語データと記憶手段に保持された過去の単語データとを１つの単語データに合成する単語管理手段とをさらに有し、異なる時間属性を持つ複数の単語データを管理することにより、専門性や話題の時間的な変化に対して良好に追従可能な形態で個人プロファイルを生成することができる。

【００２１】

【発明の実施の形態】本発明の一実施形態に係る個人プロファイル管理装置を、図面を参照して説明する。図１には、本実施形態に係る個人プロファイル管理装置の全体構成を示してあり、当該個人プロファイル管理装置は、複数のクライアントシステム１と１つのサーバシステム３とを、これらの間の通信を行うネットワーク４で接続して構成されている。

【００２２】各クライアントシステム１は、送信情報獲

得部10、テキスト記憶部11、単語分割／頻度計算部12、高頻度単語記憶部13、高頻度単語管理部14、前期高頻度単語記憶部15、および、通信インタフェース部16を備えている。これらクライアントシステム1はユーザ毎に存在し、通信インタフェース部16とコンピュータネットワーク4を通してサーバシステム3と互いに通信を行う。また、各クライアントシステム1には、電子メールシステム20、個人のｗｅｂクライアント21、個人のｗｅｂサーバ22が備えられている。

【0023】サーバシステム3は、通信インタフェース部31、高頻度単語獲得部32、個人－高頻度単語テーブル記憶部33、単語出現人数計算部34、高頻度単語－出現人数テーブル記憶部35、個人別単語顕現度計算部36、および、個人プロファイル記憶部37を備えている。なお、クライアントシステム1およびサーバシステム3はコンピュータハードウエア資源を用いて所定のプログラムを実行することにより構成されている。

【0024】各クライアントシステム1の各機能手段はそれぞれ下記のように動作して、通信インタフェース部16からネットワーク4を介してサーバシステム3へ単語データを送信する。まず、送信情報獲得部10は、電子メールシステム20、個人のｗｅｂクライアント21、個人のｗｅｂサーバ22などが、他者の管理下にあるクライアントシステム1に情報を送信したこと、および、当該クライアントシステム1において他者の管理下にあるクライアントシステム1が情報を入手することができる状態にしたことを検出し、これら送信される情報からテキスト部分を獲得する。例えば、電子メールシステム20とｗｅｂクライアント21の場合、SMTP,HTTP,FTPなどのプロトコルにしたがって、他者の管理下にあるクライアントシステム1に情報を送信したことを検出する。また、個人のｗｅｂサーバ22の場合、他者のアクセスを許す特定のディレクトリなどに情報を記憶したことを検出し、記憶された情報からテキスト部分を獲得する。

【0025】次に、テキスト記憶部11は、送信情報獲得部10が獲得したテキストをそれまでの記憶内容に追加して、ファイル単位で記憶する。また、後述するように高頻度単語管理部14から単語分割／頻度計算部12の処理の終了を通知された場合に、テキスト記憶部11は記憶している全テキスト内容をクリアにして、不必要な記憶容量の圧迫を回避する。

【0026】次に、単語分割／頻度計算部12は、例えば公知の形態素解析技術を用いて、テキスト記憶部11に記憶されたテキストから単語を抜き出し、これら単語の総数を計数する。なお、このとき、プロファイルとして不適な単語を登録した不用語テーブルを用意しておき、不用語テーブル中に存在する単語については、以降の処理を行わないようにして処理負担を軽減することも可能である。そして、単語分割／頻度計算部12は、抜き出した各単語に重複があれば、同じ単語がいくつ存在するかを計数し、各単語とその出現回数とを組みとした単語データを作成し、出現回数が所定の条件を満たすデータに限り、高頻度単語記憶部13に記憶する。ここに、所定の条件には、例えば、出現回数の大きい単語から上位W個、出現回数がX以上の単語、あるいは、出現回数が大きい単語から上位（Ｙ÷総単語数）個、出現回数が（Ｚ÷総単語数）以上の単語、などの条件を用いることができる。なお、この所定の条件は、他者への内容開示の制限、処理の効率化、および、記憶容量の有効利用のために、出現回数の大きい単語だけをサンプリングするための条件であればこれらに限らない。

【0027】図2には、高頻度単語記憶部13に記憶される高頻度単語データの一例を示してある。図示のように、単語データは、抜き出された単語の総数「１９０」と、各単語に対応付けたその出現頻度が一覧として含まれている。なお、この単語の総数には単語の重複が含まれている。例えば、抜き出された単語「意見」は、処理対象のテキスト中に延べ１１個存在することが表されている。

【0028】高頻度単語管理部14は、高頻度単語記憶部13とテキスト記憶部11のクリアおよび高頻度単語データの更新（手続きＡ）を行い、さらにまた、高頻度単語記憶部13と前期高頻度単語記憶部15の内容を統合して、通信インタフェース部16を通して高頻度単語データをサーバシステム3に通知（手続きＢ）する。なお、前期高頻度単語記憶部15には、高頻度単語記憶部13のクリアに際して当該記憶内容（クリア対象の高頻度単語データ）が記憶され、その結果、高頻度単語記憶部13が記憶している高頻度単語データに対して過去に生成された高頻度単語データが記憶されている。

【0029】高頻度単語管理部14が行う手続きＡは図3に示す処理手順で実行され、当該手続きＡはテキストが或る量以上に貯まったところで繰り返し実行される。まず、テキスト記憶部11の記憶されているテキスト量を調べ、テキスト量が閾値を越えたかどうかを判断する（ステップＳ1）。そして、テキスト量が閾値を越えた場合には、当該テキスト中から単語の抜き出しと抜き出された単語数の計数を行い（ステップＳ2）、ｎ番目の単語がｗａ［ｎ］、ｗａ［ｎ］の出現回数がｗｆａ［ｎ］の組みからなるデータ｛ｗａ［ｎ］，ｗｆａ［ｎ］｝を作成する（ステップＳ3）。

【0030】次いで、高頻度単語記憶部13の記憶内容を読み出し、ｍ番目の単語がｗｂ［ｍ］、ｗｂ［ｍ］の出現回数がｗｆｂ［ｍ］の組みからなるデータ｛ｗｂ［ｍ］，ｗｆｂ［ｍ］｝を作成する（ステップＳ4）。そして、｛ｗａ［ｎ］，ｗｆａ［ｎ］｝と｛ｗｂ［ｍ］，ｗｆｂ［ｍ］｝との論理和｛ｗ［ｋ］，ｗｆ［ｋ］｝とし（ステップＳ5）、この｛ｗ［ｋ］，ｗｆ［ｋ］｝の要素を調べて、ｗ［ｉ］＝ｗ［ｊ］の場合に

は、ｗｆ［ｉ］＋ｗｆ［ｊ］を新たなｗｆ［ｉ］として重複をまとめた後に、一方のｗ［ｊ］とｗｆ［ｊ］の組を削除する（ステップＳ６）。そして、高頻度単語記憶部１３の単語総数を抜き出した単語数だけ増加させ（ステップＳ７）、出現頻度ｗｆ［ｉ］が所定の閾値（Ｘ）以上の組だけを高頻度単語記憶部１３に上書きし（ステップＳ８）、テキスト記憶部１１の記憶内容をクリアする（ステップＳ９）。なお、上記の一連の処理は、テキスト記憶部１１が記憶するテキスト量が閾値を超えると繰り返し行われる。

【００３１】また、高頻度単語管理部１４が行う手続きＢは図４に示す処理手順で実行され、当該手続きＢは一定の時間間隔で実行される。まず、手続きＢが前回起動されてからの時間を調べ、経過時間が一定値を越えたかどうかを判断する（ステップＳ１１）。そして、経過時間が一定値を越えた場合には、高頻度単語記憶部１３の単語総数をＷＣ１とするとともに、前期高頻度単語記憶部１５の単語総数をＷＣ２とし（ステップＳ１２）、さらに、高頻度単語記憶部１３の単語と頻度の組みを｛ｗ１［ｎ］，ｗｆ１［ｎ］｝とするとともに、前期高頻度単語記憶部１５の単語と頻度の組みを｛ｗ２［ｍ］，ｗｆ２［ｍ］｝とする（ステップＳ１３）。

【００３２】次いで、（ＷＣ１＋ＷＣ２）÷２をＷＣとして、すべてのｎとｍについて、ｗｆ１［ｎ］×ＷＣ１÷ＷＣを新しいｗｆ１［ｎ］とするとともに、ｗｆ２［ｍ］×ＷＣ２÷ＷＣを新しいｗｆ２［ｍ］とし（ステップＳ１４）、さらに、｛ｗ１［ｎ］，ｗｆ１［ｎ］｝と｛ｗ２［ｍ］，ｗｆ２［ｍ］｝との論理和を｛ｗ［ｋ］，ｗｆ［ｋ］｝とし（ステップＳ１５）、この｛ｗ［ｋ］，ｗｆ［ｋ］｝の要素を調べて、ｗ［ｉ］＝ｗ［ｊ］の場合には、ｗｆ［ｉ］＋ｗｆ［ｊ］を新たなｗｆ［ｉ］として重複をまとめた後に、一方のｗ［ｊ］とｗｆ［ｊ］の組を削除する（ステップＳ１６）。そして、出現頻度ｗｆ［ｋ］が所定の閾値（Ｘ）以上の組だけを高頻度単語データとして通信インタフェース部１６を通してサーバシステム３に通知する（ステップＳ１７）。

【００３３】図５には、前期高頻度単語記憶部１５に記憶されている単語データの一例を示してあり、この過去の単語データでは、単語の総数が「２２１」で、例えば「君」という単語は延べ９個存在することを表している。また、図６には、図２に示した高頻度単語データと図５に示した前期高頻度単語データに対して、上記の手続きＢを行った後に得られた高頻度単語データの一例を示してある。手続きＢを行って得られた高頻度単語データでは、単語総数は平均値の「２０５．５」、例えば、「検討」という単語は６×２２１÷２０５．５＝６．５の頻度となっている。

【００３４】なお、上記の手続きＡにおいて出現頻度ｗｆ［ｋ］が所定の閾値（Ｘ）以上の組だけを選択するようにし、また、上記の手続きＢにおいて出現頻度ｗｆ［ｋ］が所定の閾値（Ｘ）以上の組だけを選択するようにしたが、本発明では、このような選択条件に限らず、他者への内容開示の制限、処理の効率化、および、記憶容量の有効利用のために、出現頻度ｗｆ［ｋ］の大きいものだけをサンプリングするための条件であれば、種々な条件を設定することができる。例えば、出現頻度ｗｆ［ｋ］が大きいものから上位Ｗ個を選択する、出現頻度ｗｆ［ｋ］が大きいものから上位（Ｙ÷総単語数）個を選択するなどの条件を用いることができる。また、上記の手続きＡはテキスト量に応じて開始され、上記の手続きＢは一定の時間間隔で開始されるようにしたが、例えば、通信インタフェース部を通してサーバシステム３から手続きの起動の指示を受け取ることによって一連の処理を開始するようにすることもできる。

【００３５】上記のようにして、各クライアントシステム１で生成された高頻度単語データがサーバーシステム３に対して送信され、サーバシステム３の各機能手段はそれぞれ下記のように動作して、これら高頻度単語データに基づいて個人プロファイルを生成する。まず、高頻度単語獲得部３２は、複数のクライアントシステム１から通信インタフェース部３１を介して通知される高頻度単語データを収集し、これら高頻度単語データを個人毎に集計して、個人－高頻度単語テーブルを生成して記憶部３３に記憶させる。なお、或る個人が複数のクライアントシステム１を利用する可能性がある場合には、各個人の識別子あるいは各クライアントシステム１の識別子から特定の個人を同定し、同一の個人の高頻度単語データであれば、前記手続きＡと同様の方法で高頻度単語データを統合するようにすればよい。

【００３６】図７には、高頻度単語獲得部３２が収集した高頻度単語データから生成した個人－高頻度単語テーブルの一例を示してある。この個人－高頻度単語テーブルは、Ａさん、Ｂさん、Ｃさん、および、Ｄさんの４人分の例であるが、例えば「私」や「問題」といった単語のように、多くの個人のテキスト中に高頻度で出現するために、必ずしもＡさんやＢさんなどといった特定の個人を特徴付けることにはならない単語も上位に存在している。したがって、単語とその出現頻度との関係だけでは、個々人の特徴を十分には判別することができない。

【００３７】単語出現人数計算部３４は、単語毎に、その単語を個人－高頻度単語テーブル中に含む個人の人数を計算し、高頻度単語－出現人数テーブルを生成して記憶部３５に記憶させる。図８には、高頻度単語－出現人数テーブルの一例を示してあり、この高頻度単語－出現人数テーブルは、図７に示した４人を含む６人の個人－高頻度単語テーブルから生成されている。例えば、「私」や「問題」は６人中３人の個人－高頻度単語テーブル中に含まれており、多くの個人のテキスト中に高頻度で出現するために顕現度が低く、必ずしも特定の個人

11

を特徴付けることにはならない単語となっている。

【0038】個人別単語顕現度計算部36は、個人－高頻度単語テーブルと高頻度単語－出現人数テーブルの内容から、個人プロファイルを生成して記憶部37に記憶させる。この個人別単語顕現度計算部36での処理は次のように行われるが、以下の説明では、対象となる人数をP、個人の個人－高頻度単語テーブル中のn番目（全＊

12

＊部でN個）の個人の単語と頻度の組みを｛w［n］，wf［n］｝、高頻度単語－出現人数テーブル中のj番目の単語と人数の組みを｛wr［j］，pf［j］｝で表現する。このとき、単語w［n］の顕現度p［n］を式（2）で計算する。

【0039】

【数2】

$$p[n] = wf[n] \times \ln\left(P \div pf[j]\ |_{w[n]=wr[j]}\right)$$

ここに、$pf[j]\ |_{w[n]=wr[j]}$ は、w［n］とwr［j］とが同一であるjについてのpf［j］を表すものとする。

$$\cdots\cdots (2)$$

【0040】例えば、Bさんの場合、顕現度は以下の手順で計算する。n＝1のとき、個人－高頻度単語テーブルではw［1］＝「意見」の出現頻度wf［1］は13である。一方、「意見」は高頻度単語－出現人数テーブル中では12番目に相当し、その出現人数pf［12］は1である。したがって、式（2）によって、顕現度p［1］＝13×ln（6÷1）＝23．3となる。また、n＝2のとき、個人－高頻度単語テーブルではw［2］＝「君」の出現頻度wf［2］は9である。一方、「君」は高頻度単語－出現人数テーブル中では1番目に相当し、その出現人数pf［1］は3である。したがって、式（2）によって、顕現度p［1］＝9×ln（6÷3）＝6．23となる。以下同様にして、N番目までの単語について顕現度p［n］を求める。このように、各個人毎の各単語の出現頻度を全人数に対する当該単語を用いた人数の対数で補正（すなわち、使用した人※数が少なければ大きな値で出現頻度を補正）することにより、各個人における各単語が当該個人の特徴をどの程度表しているかを定量化することができる。

【0041】なお、この顕現度p［n］の値を他者と比較するために、さらに正規化を行うことも可能である。その場合、個人内での他の単語の顕現度の自乗和の平方を求めて、各単語の顕現度との比率を正規化顕現度とする。正規化顕現度pn［n］は式（3）で計算する。例えば、「意見」の正規化顕現度はpn［1］は、23．3÷51．60＝0．45である。図8には、A～Dの4人について正規化顕現度を求め、その正規化顕現度の大きい順に単語を並べたテーブルの一例を示してあり、このテーブルが個人プロファイル記憶部37に記憶される個人プロファイルである。

【0042】

【数3】

$$pn[n] = p[n] \div \left(\sum_{k=1}^{N} p[k]^{s}\right)^{0.5} \qquad \cdots\cdots (3)$$

【0043】なお、上記の例では、顕現度pn［n］の計算において自然対数lnを用いたが、例えば任意の正数aを底とする対数$\log_a$を用いるように変更してもよく、要は、全体の人数Pに対して、或る単語を使っている人数が多い場合に、顕現度pn［n］の値を小さくすることができれば、これらの関数に限定するものではない。

【0044】図10には、本発明の他の一実施形態に係る個人プロファイル管理装置の全体構成を示してあり、本実施形態は、クライアントシステム1が自ら送信するテキストに加えて受信したテキストからも単語を抽出して個人プロファイルを作成するものである。本実施形態の個人プロファイル管理装置は、複数のクライアントシステム1と、これらの間の通信を行うネットワーク4で構成されている。なお、図示は省略してあるが、前日した実施形態と同様なサーバシステム3も当該ネットワーク4に接続されている。ここで、以下の説明において、前述した実施形態と同様な機能手段については同一符号を付して重複する説明を割愛する。

【0045】少なくとも1つのクライアントシステム1には、送信情報獲得部10、受信情報獲得部17、テキスト記憶部11、単語分割／頻度計算部12、高頻度単語記憶部13、高頻度単語管理部14、前期高頻度単語記憶部15、個人－高頻度単語テーブル記憶部33、単語単語出現人数計算部34、高頻度単語－出現人数テーブル記憶部35、個人別単語顕現度計算部36、個人プロファイル記憶部37、および、通信インタフェース部31が備えられている。すなわち、本実施形態では、前述した実施形態（図1）に較べて、サーバシステム3側に設けられていた各機能手段32～37を高頻度単語獲得部32を除いてクライアントシステム1側に設け、また、当該クライアントシステムに受信情報獲得部17を新たに設けた構成となっている。

【0046】この受信情報獲得部17は、電子メールシステム20、個人のwebクライアント21、webサーバ22などが、他者の管理下にあるクライアントシステム1から情報を受信したことを検出し、受信された情報からテキスト部分を獲得する。例えば、電子メールシ

ステム２０とｗｅｂクライアント２１の場合、SMTP,HTTP,FTPなどのプロトコルにしたがって、他者の管理下にあるクライアントシステム１から情報を受信したことを検出する。

【００４７】ここで、テキスト記憶部１１は、前述した実施形態と同様に送信情報獲得部１０が獲得したテキストの記憶処理を行うが、これに加えて、受信情報獲得部１７が獲得したテキストを、それまでの記憶内容に追加して、ファイル単位で個人毎に記憶する処理も行う。また、単語分割／頻度計算部１２は、テキストから単語を抜き出して単語の総数を計数するといった前述の実施形態と同様の処理を行うが、本実施形態では、単語とその出現回数とを組みとした単語データを、１人の個人についてだけでなく、当該クライアントシステム１と通信する複数の個人別に作成する。

【００４８】また、高頻度単語管理部１４は、高頻度単語記憶部１３とテキスト記憶部１１のクリアおよび高頻度単語データの更新（手続きＣ）を行い、また、高頻度単語記憶部１３と前期高頻度単語記憶部１５の内容を統合して、通信インタフェース部３１を通して高頻度単語データを図外のサーバシステム３に通知（手続きＤ）する。この手続きＣでは、一定の時間間隔で以下の一連の処理を行う。まず、個人識別用の変数を初期化し、現在の個人識別用の変数が指し示す個人について、前述した手続きＡ（図３）を行う。そして、個人識別用の変数をインクリメントし、もし手続きＡが終了していない個人があれば、その個人について手続きＡを実行する。もし、すべての対象となる個人について手続きＡが終了したならば、再度、一定の時間間隔が経過するのを待って処理を繰り返し行う。

【００４９】また、手続きＤは手続きＢ’を用いてなされ、まず、手続きＢ’を説明すると、手続きＢ’は前述した手続きＢ（図４）において、「ｗｆ［ｋ］がＸ以上のデータを、通信インタフェース部を通してサーバシステムに通知する（ステップＳ１７）」処理のみを、新しく「ｗｆ［ｋ］がＸ以上のデータを、個人－高頻度単語テーブル記憶部に記憶する」に変更した手続きである。そして、手続きＤでは、まず、個人識別用の変数を初期化し、現在の個人識別用の変数が指し示す個人について、手続きＢ’を行う。そして、個人識別用の変数をインクリメントし、もし手続きＢ’が終了していない個人があれば、その個人について手続きＢ’を実行する。もし、すべての対象となる個人について手続きＢ’が終了したならば、再度、一定の時間間隔が経過するのを待って処理を繰り返し行う。

【００５０】

【発明の効果】以上説明したように、本発明によれば、テキストを各個人が利用するクライアントシステムで処理して抽出された単語に基づいて個人プロファイルを作成するようにしたため、テキストを意図しない他者が処理することがなく、電子メールテキスト等のように内容の他者への開示を制限したいという要求を満すことができる。さらに、他者に対する各個人の単語の相対的な顕現度を加味して個人プロファイルを定めるため、各個人の特徴をよく表した個人プロファイルを生成することができる。さらに、異なる時間属性を持つ複数の高頻度単語データを管理するようにしたため、専門性や話題の時間的な変化に対して良好に追従可能な形態で個人プロファイルを管理することができる。

【図面の簡単な説明】

【図１】　本発明の一実施形態に係る個人プロファイル管理装置の構成図である。

【図２】　高頻度単語データの一例を示す図である。

【図３】　高頻度単語管理部の手続きＡの処理手順を示すフローチャートである。

【図４】　高頻度単語管理部の手続きＢの処理手順を示すフローチャートである。

【図５】　前期高頻度単語データの一例を示す図である。

【図６】　手続きＢを施した後の高頻度単語データの一例を示す図である。

【図７】　個人－高頻度単語テーブルの一例を示す図である。

【図８】　高頻度単語－出現人数テーブルの一例を示す図である。

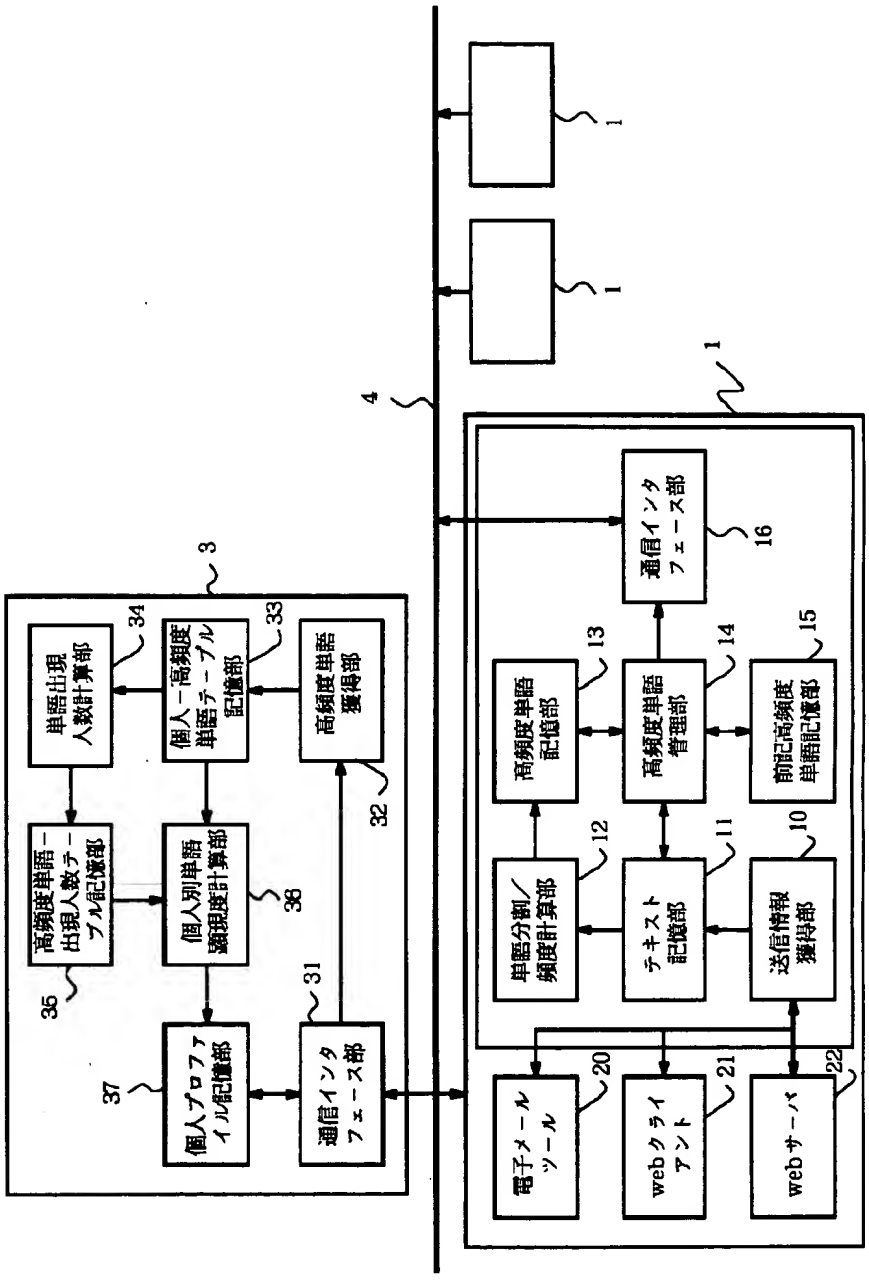
【図９】　個人プロファイルの一例を示す図である。

【図１０】　本発明の他の一実施形態に係る個人プロファイル管理装置の構成図である。

【符号の説明】

１・・・クライアントシステム、　３・・・サーバシステム、４・・・ネットワーク、　１０・・・送信情報獲得部、１１・・・テキスト記憶部、　１２・・・単語分割／頻度計算部、１３・・・高頻度単語記憶部、　１４・・・高頻度単語管理部、１５・・・前期高頻度単語記憶部、　１６・・・通信インタフェース部、３１・・・通信インタフェース部、　３２・・・高頻度単語獲得部、３３・・・個人－高頻度単語テーブル記憶部、３４・・・単語出現人数計算部、３５・・・高頻度単語－出現人数テーブル記憶部、３６・・・個人別単語顕現度計算部、　３７・・・個人プロファイル記憶部、

【図1】


1
1
4
1
3
単語出現人数計算部
34
個人−高頻度単語テーブル記憶部
33
高頻度単語獲得部
32
高頻度単語−出現人数テーブル記憶部
35
個人別単語頻度計算部
36
個人プロファイル記憶部
37
通信インタフェース部
31
通信インタフェース部
16
高頻度単語記憶部
13
高頻度単語管理部
14
前記高頻度単語記憶部
15
単語分割／頻度計算部
12
テキスト記憶部
11
送信情報獲得部
10
電子メールツール
20
webクライアント
21
webサーバ
22

【図２】

単語	頻度	単語総数	190
意見	11		
交換	6		
SG	6		
研究	5		
交流	4		
データ	4		
85G	3		
コメント	3		
1	3		
こと	3		

高頻度単語データの一例

【図５】

単語	頻度	単語総数	221
君	9		
検討	6		
85G	5		
こと	5		
会	5		
吉田	5		
変更	5		
たか	4		
グループ	3		
研究	3		

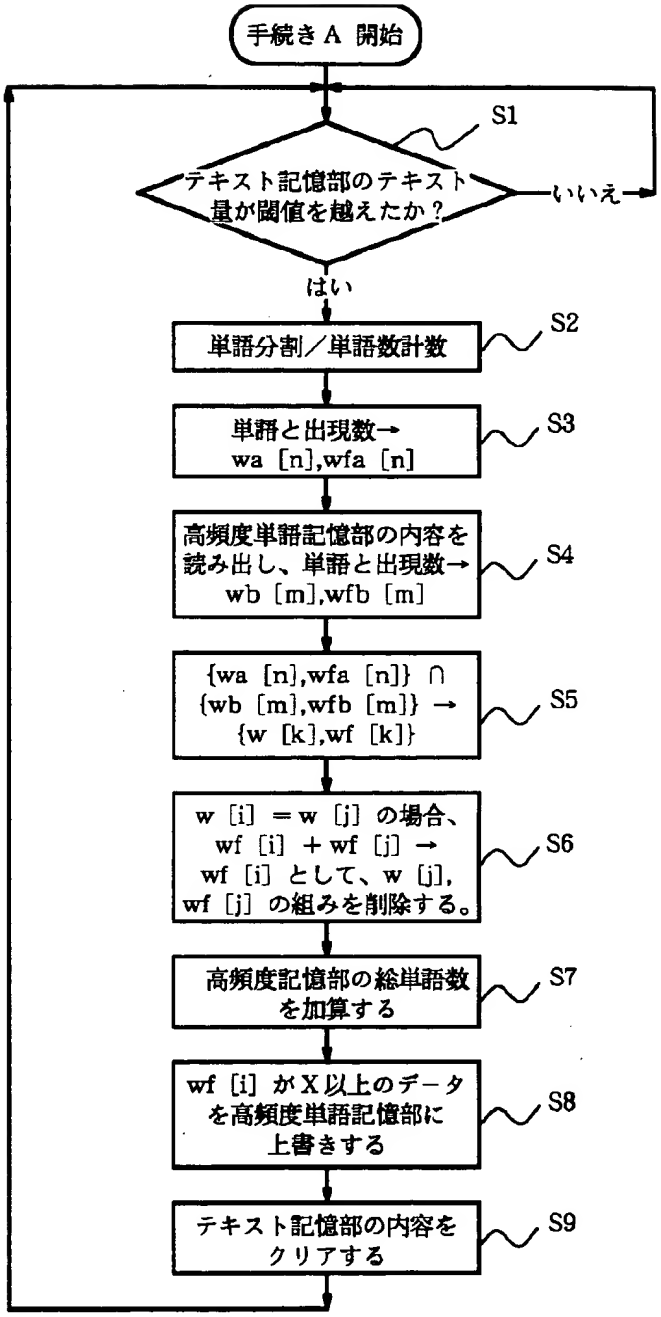
前記高頻度単語データの一例

【図６】

単語	頻度	単語総数	205.5
意見	10.2		
君	9.7		
検討	6.5		
交換	5.5		
SG	5.5		
85G	5.4		
こと	5.4		
会	5.4		
吉田	5.4		
変更	5.4		

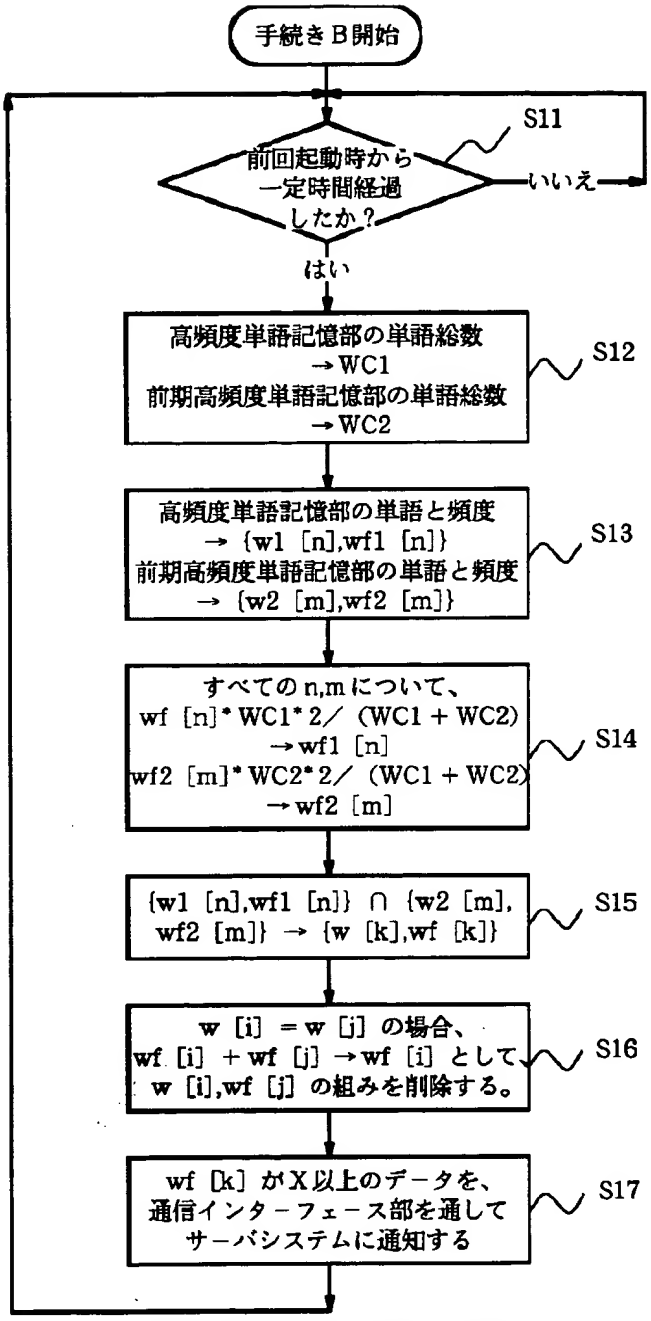
手続きＢ後の高頻度単語データ

【図３】



高頻度単語管理部の手続きＡの動作

【図４】



高頻度単語管理部の手続きＢの動作

【図８】

単語	出現人数
君	3
山下	3
私	3
問題	3
こと	2
サークル	2
活動	2
実習	2
水曜日	2
SG	1
たか	1
連絡	1
会	1
記憶	1
吉田	1
研究	1
交換	1
変更	1

高頻度単語－出現人数テーブルの一例

【図７】

A		B		C		D	
こと	5	意見	13	登録	4	NX	3
活動	5	君	9	メール	3	運動	3
水曜日	3	85G	8	文字	3	1	2
サークル	2	こと	8	Common	2	85G	2
記憶	2	検討	8	Property	2	君	2
山下	2	研究	8	カテゴリ	2	参加	2
私	2	SG	6	化け	2	山下	2
実習	2	吉田	6	件	2	社	2
問題	2	交換	6	今回	2	金	2
7：00	1	たか	5	設定	2	7	1
30	1	会	5	こ	1	97	1
85	1	変更	5	コード	1	Forum	1
PM5	1	2	4	ところ	1	New	1
QC	1	グループ	4	パソコン	1	Way	1
アイデア	1	データ	4	一	1	Work	1
グループ	1	交流	4	下	1	メッセ	1
ケア	1	参加	4	何度	1	何	1
これ	1	中	4	確認	1	火	1
ごろ	1	問題	4	高橋	1	開催	1
それ	1	1	3	再度	1	希望	1

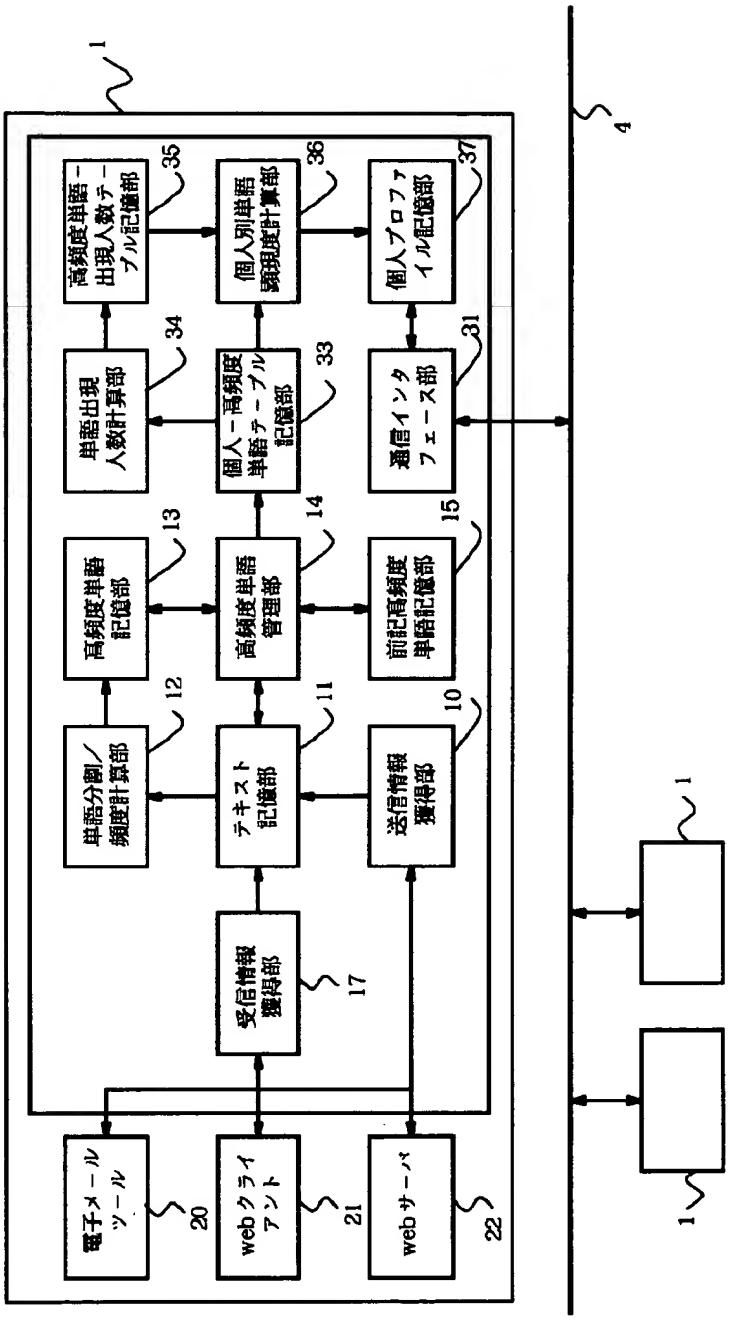
個人－高頻度単語テーブルの一例

【図９】

A		B		C		D	
単語	優先度	単語	優先度	単語	優先度	単語	優先度
こと	0.55	意見	0.45	件	0.44	NX	0.55
活動	0.55	研究	0.28	設定	0.44	運動	0.55
記憶	0.36	SG	0.21	文字	0.40	社	0.37
水曜日	0.33	吉田	0.21	登録	0.34	金	0.37
サークル	0.22	交換	0.21	カテゴリ	0.27	1	0.22
実習	0.22	たか	0.17	化け	0.27	85G	0.14
山下	0.14	会	0.17	今回	0.27	君	0.14
私	0.14	変更	0.17	メール	0.25	参加	0.14
問題	0.14	こと	0.17	Common	0.17	山下	0.14

個人プロファイルの一例

【図10】


1
4
高頻度単語－出現人数テーブル記憶部
35
個人別単語頻現度計算部
36
個人プロファイル記憶部
37
単語出現人数計算部
34
個人－高頻度単語テーブル記憶部
33
通信インタフェース部
31
高頻度単語記憶部
13
高頻度単語管理部
14
前記高頻度単語記憶部
15
単語分割／頻度計算部
12
テキスト記憶部
11
送信情報獲得部
10
受信情報獲得部
17
電子メールツール
20
webクライアント
21
webサーバ
22
1
1